1 [1,5 puntos] Describa en qué consiste la jerarquía de memoria, cuáles son sus componentes principales, en qué principios basa su funcionamiento y describa brevemente el objetivo y características de las políticas de ubicación y reemplazo.

SOLUCIÓN

La jerarquía de memoria (JM) es la organización de la memoria del computador por niveles, de modo que se disponga de una memoria rápida aunque costosa próxima al procesador y de memoria de gran capacidad aunque sea lenta en los niveles más alejados de la CPU.

Los componentes principales son las memorias caché (normalmente de 1 a 3 niveles de caché), la memoria principal y el dispositivo de almacenamiento secundario como soporte de la memoria virtual.

La JM se basa en los principios de localidad de referencias espacial y temporal que presentan en distinta medida los programas reales.

La política de ubicación determina en qué posición/posiciones de la memoria de un cierto nivel de la JM puede situarse la información que se lleva desde el nivel siguiente. Puede definirse una política directa, asociativa o asociativa por conjuntos.

La política de reemplazo se utiliza para seleccionar qué información se desaloja de un cierto nivel de la JM cuando se va a incorporar una nueva información. Se utilizan principalmente la política aleatoria, FIFO, LRU y variaciones sobre ellas.

2 [1,5 puntos] Responda razonadamente si son ciertas o no las siguientes afirmaciones sobre el sistema de memoria:

a) La memoria entrelazada es una técnica software para reducir la tasa de fallos de una memoria caché.

b) En un sistema de memoria que incluye memoria caché, el tiempo invertido en un acceso de lectura es independiente de la política de escritura de la memoria caché.

c) El sistema de memoria virtual permite proteger información privada de procesos y, a la vez, facilitar la compartición de información entre esos mismos u otros procesos.

SOLUCIÓN

a) Falsa. Se trata de una organización de la memoria principal en módulos de modo que se facilite el acceso en paralelo a varios de ellos, permitiendo así mejorar el ancho de banda de este componente de la jerarquía de memoria. El entrelazado puede organizarse con un esquema simple (acceso simultáneo a los módulos) o complejo (acceso independiente, ciclo partido). Según la distribución de direcciones, el entrelazado puede ser inferior (direcciones consecutivas se alojan en diferentes módulos de memoria) o superior (diferentes regiones de memoria se alojan en distintos módulos).

b) Falsa. El tiempo invertido en un acceso de lectura cuando se da un fallo de caché depende, entre otros factores, del estado del bloque que se reemplaza: si se encuentra modificado, el acceso tendrá una duración considerablemente mayor. Por otro lado, si se utiliza una política de escritura inmediata (*write through*), nunca se tendrán que reemplazar bloques de caché modificados, al contrario de lo que ocurre si la política de escritura es aplazada (*copy back*).

c) Verdadero. El sistema de memoria virtual no solo permite eliminar algunas limitaciones en cuanto al tamaño de la memoria disponible para un proceso, sino que, también independiza los espacios de memoria de los distintos procesos, a la vez que facilita mecanismos de compartición entre ellos.

PROBLEMA (responda en otra hoja)

3 Considere un computador dotado de memoria virtual paginada y un único nivel de memoria caché que presenta las siguientes características:

- Palabras de 32 bits, con buses de direcciones y datos de 32 bits.
- Direcciones virtuales de 46 bits con 3 niveles de tablas de páginas y TLB.
- Páginas de 8 KBytes. Cada entrada de cualquiera de las tablas de páginas ocupa una palabra.
- Cada una de las tablas de páginas de cualquier nivel ocupa 1 página.
- Memoria caché de 1MB asociativa por conjuntos de 8 bloques con:
  - Bloques de 64 Bytes
  - Escritura aplazada (*copy back with allocation*, CBWA) y lectura "out of order fetch" (OOF).

a) [2 puntos] Describa los campos en que se descompone una dirección según es interpretada por el sistema de memoria virtual y por la memoria caché, especificando en ambos casos el tamaño en bits de cada campo. Justifique si es posible realizar simultáneamente el acceso a a la memoria caché y a la TLB.

Considere además, para el resto del problema, los siguientes datos y estimaciones:

- No se producen fallos de página.
- El 20 % de los accesos a memoria son de escritura.
- Tiempos de acceso. TLB: 5 ns. Memoria caché: 5 ns.
- Tiempo de acceso de la memoria principal:
  - Lectura/escritura de una palabra: 50 ns.
  - Lectura/escritura de un bloque de 16 palabras: 200 ns.
- Tasas de acierto estimadas. TLB: 96 %. Memoria caché: 90 %.
- La probabilidad de que el bloque de caché a reemplazar esté modificado es del 10 %.

b) [1 punto] Calcule el tiempo medio de traducción invertido en cada acceso a memoria.

c) [2 puntos] Calcule el tiempo medio de acceso, distinguiendo entre lecturas y escrituras.

d) [1 punto] Calcule el tiempo medio de ocupación en el sistema y las condiciones indicados.

e) [1 punto] Justifique por qué los computadores reales tienen desdoblada la memoria caché en una específica para instrucciones y otra para datos.

SOLUCIÓN

a) Las páginas son de 8 KBytes, es decir 2^{13} Bytes, por lo que se necesitan 13 bits para identificar cualquier byte dentro de una página. Por otra parte, cualquier tabla de páginas ocupa 1 página, 2^{13} Bytes, y cualquier entrada ocupa una palabra, 2^2 Bytes, con lo que una tabla de páginas tiene $2^{13}/2^2 = 2^{11}$ entradas y se necesitan 11 bits para acceder a dicha entrada. A partir de estas indicaciones se define el formato siguiente para la dirección virtual:

Zona	Región	Página	Byte
11 bits	11 bits	11 bits	13 bits

El tamaño de la memoria caché es de 1 MByte, 2^{20} Bytes, e interpreta direcciones físicas que constan de 32 bits. Los campos en que interpretará estas direcciones son la etiqueta, el conjunto y el desplazamiento dentro del bloque. El desplazamiento viene determinado por el tamaño del bloque, que al ser de 2^6 Bytes requerirá 6 bits. El conjunto se determina a partir del número de bloques de caché y la relación entre bloques y conjuntos:

$$N\_Bloques = 2^{20}/2^6 = 2^{14}\ Bloques. \qquad N\_Conjuntos = N\_Bloques/8 = 2^{14}/2^3 = 2^{11}\ Conjuntos$$

El resto, hasta completar los 32 bits, corresponden a la etiqueta, por lo que la interpretación es la siguiente:

Etiqueta	Conjunto	Byte
15 bits	11 bits	6 bits

No se puede solapar el acceso a TLB con el acceso a memoria caché puesto que se requeriría un mínimo de 11+6=17 bits de la dirección física antes de realizar la traducción y solo se dispone de los 13 que se obtienen del desplazamiento dentro de la página virtual.

b) Para calcular el tiempo de traducción basta con saber que se dispone de una TLB cuyo tiempo de acceso es de 5 ns, que hay tres niveles de tablas de páginas y que se considera una tasa de aciertos de 0,96 para la TLB:

$$\overline{t}_{trad} = t_{TLB} + (1 - Hr_{TLB}) \times N_{NivelesTP} \times t_{Mp}$$

$$\overline{t}_{trad} = 5\ ns + 0,04 \times 3 \times 50\ ns = 11,0\ ns$$

c) El tiempo medio de acceso en caso de lecturas dependerá de los tiempos y probabilidad de encontrar en caché la información que se busca (5 ns y 0,9 respectivamente), de la probabilidad de que en caso de fallo el bloque a reemplazar se encuentre modificado (0,1 según el enunciado) y del tiempo de escritura de un bloque completo en memoria principal (200 ns), por lo que será:

$$\overline{t}_{accL} = \overline{t}_{trad} + t_{Mca} + (1 - Hr_{Mca}) \times (t_{RD1pal} + P_{BLmod} \times t_{WRbloque})$$

$$\overline{t}_{accL} = 11\ ns + 5\ ns + (1 - 0,9) \times (50\ ns + 0,1 \times 200\ ns) = 23,0\ ns$$

Para hacer este cálculo se ha considerado que la política de lectura utilizada es *out of order fetch*, por lo que basta con leer la primera palabra de un bloque para contabilizar el tiempo de acceso. En el caso de las escrituras se puede considerar que en primer lugar se comprueba si el bloque a reemplazar está modificado, en cuyo caso se escribe en memoria principal. A continuación se escribe la palabra que ha producido el fallo en memoria principal y finalmente se lleva el bloque de memoria principal a caché. De este modo, el tiempo medio de acceso en caso de escritura será el mismo que el calculado para las lecturas:

$$\overline{t}_{accE} = \overline{t}_{trad} + t_{Mca} + (1 - Hr_{Mca}) \times (P_{BLmod} \times t_{WRbloque} + t_{WR1pal})$$

$$\overline{t}_{accE} = 11\ ns + 5\ ns + (1 - 0,9) \times (0,1 \times 200\ ns + 50\ ns) = 23,0\ ns$$

Así pues, para el cálculo del tiempo medio de acceso se sumarán los tiempos de lectura y escritura ponderados (en realidad, puesto que son idénticos, la ponderación no sería necesaria):

$$\overline{t}_{acc} = P_{RD} \times \overline{t}_{accL} + P_{WR} \times \overline{t}_{accE}$$

$$\overline{t}_{acc} = 0,8 \times 23\ ns + 0,2 \times 23\ ns = 23,0\ ns$$

d) La diferencia entre el tiempo medio de ocupación y el de acceso calculado en el apartado anterior se da únicamente en los accesos con fallo de caché. De este modo, en los accesos de lectura con fallo hay que considerar el tiempo de lectura de un bloque completo, no solo de su primera palabra. En los accesos de escritura hay que añadir el tiempo de lectura del bloque modificado una vez que se ha completado la escritura de la correspondiente palabra.

$$\overline{t}_{ocupL} = \overline{t}_{trad} + t_{Mca} + (1 - Hr_{Mca}) \times (t_{RDbloque} + P_{BLmod} \times t_{WRbloque})$$

$$\overline{t}_{ocupL} = 11\ ns + 5\ ns + (1 - 0,9) \times (200\ ns + 0,1 \times 200\ ns) = 38,0\ ns$$

$$\overline{t}_{ocupE} = \overline{t}_{trad} + t_{Mca} + (1 - Hr_{Mca}) \times (P_{BLmod} \times t_{WRbloque} + t_{WR1pal} + t_{RDbloque})$$

$$\overline{t}_{ocupE} = 11\ ns + 5\ ns + (1 - 0,9) \times (0,1 \times 200\ ns + 50\ ns + 200\ ns) = 43,0\ ns$$

El tiempo medio de ocupación será entonces:

$$\overline{t}_{ocup} = P_{RD} \times \overline{t}_{ocupL} + P_{WR} \times \overline{t}_{ocupE}$$

$$\overline{t}_{ocup} = 0,8 \times 38\ ns + 0,2 \times 43\ ns = 39,0\ ns$$

e) En general, desde el punto de vista del sistema, la memoria se desdobla porque así se puede acceder simultáneamente a instrucciones y a datos, de tal modo que un procesador con *pipeline* puede realizar la lectura o escritura de un dato al mismo tiempo que hace el *fetch* de una instrucción. Por otra parte, desde el punto de vista del sistema de memoria, ese desdoblamiento permite ajustar las características de las memorias caché a las necesidades de la información que manejan. Por ejemplo, en la memoria caché de instrucciones no tendrá sentido definir una política de escritura aplazada, puesto que en dicha caché no se realizan escrituras.

ARQUITECTURA DE COMPUTADORES
EXAMEN PARCIAL (29 de marzo de 2011)

Apellidos, Nombre..... **N° de Matrícula**.....

TEORIA. Responda en esta misma hoja, utilizando únicamente el espacio asignado para cada pregunta.

1 (2 puntos) Responda razonadamente si son ciertas o no las siguientes afirmaciones sobre el sistema de memoria

1.- La lectura fuera de orden (*out of order fetch*) es una técnica para reducir el tiempo de penalización en el acceso a la memoria cache.

Es cierto. Con este mecanismo la palabra que produjo el fallo se extrae primero y se reduce el tiempo de acceso. Si bien el tiempo de ocupación no varía.

2.- El tamaño de los bloques de la memoria cache influye en la tasa de aciertos y en el tiempo de penalización.

Es verdadero. El tamaño de los bloques es un factor crítico. Bloques grandes favorecen la proximidad espacial pero perjudican la proximidad temporal puesto que la memoria cache contiene menos bloques y, por lo tanto, menos localidad. Un tamaño grande de bloque influye en el tiempo de penalización puesto que, en general, se tarda más en resolver el fallo.

3.- Las utilización de tablas de página multinivel tiene como objetivo reducir la tasa de fallos en memoria principal.

No es cierto. Se utilizan tablas de página multinivel para reducir la cantidad de memoria necesaria para las páginas de traducción.

2 (1 punto) Sea un computador con palabra de 32 bits y 4 GB de memoria principal. Este computador dispone de una memoria cache unificada de 64 KB, asociativa por conjuntos de 4 bloques y el tamaño de cada bloque es de 64 bytes.

1.- Indique en qué campos se divide la dirección con la que se accede a la memoria cache y cuál es el tamaño de cada uno de ellos.

Como cada bloque tiene 64 bytes (2^6), el campo byte tiene 6 bits. Al ser la capacidad de la cache 64 KB, hay 1.024 bloques y por lo tanto 256 conjuntos (2^8). De modo que, el campo conjunto tiene 8 bits y los 18 restantes corresponden a la etiqueta.

Etiqueta	Conjunto	Byte
18 bits	8 bits	6 bits

2.- Detalle el procedimiento para localizar una palabra en esta memoria cache indicando para qué se usa cada uno de los campos de la dirección.

La dirección se interpreta según el formato anterior, de modo que con el campo "Conjunto" se selecciona el conjunto en el que puede estar el bloque de memoria principal al que pertenece dicha dirección. Para comprobar si hay acierto o fallo de cache, se compara el campo "Etiqueta" con las 4 etiquetas del directorio de la memoria cache correspondientes al "Conjunto". Si hay acierto se usa el campo "Byte" para seleccionar el byte solicitado dentro del bloque del conjunto cuya etiqueta coincide. Si no coincide ninguna etiqueta, entonces hay fallo de cache.

3 (1 punto) Este computador dispone de una memoria principal con entrelazado simple de 8 módulos con tiempo de acceso de 50 ns. Calcule los tiempos de acceso para lectura y escritura de un bloque de cache y de lectura de una palabra si el bus que conecta la memoria cache con la principal tiene un reloj de 500 MHz.

Como la memoria está organizada con entrelazado simple de 8 módulos, se puede acceder simultáneamente y extraer 8 palabras. Sin embargo, estas 8 palabras han de ser enviadas por el bus que tiene un ciclo de reloj de 2 ns. Así que el tiempo de acceso para leer una palabra es: $t_{lp} = 50ns + 2ns = 52ns$

Y para leer 8 palabras consecutivas: $t_{l8p} = 50ns + 8 \times 2ns = 66ns$

Sin embargo, un bloque de cache tiene 16 palabras consecutivas de modo que serán necesarias 2 lecturas: $t_{lbq} = 2 \times 66ns = 132ns$

Para la escritura de un bloque se deben transmitir las primeras 8 palabras por el bus, esperar el tiempo de acceso y transmitir las 8 restantes. Además, antes de realizar otra operación sobre memoria habrá que esperar el tiempo de acceso correspondiente a la escritura de las últimas 8 palabras: $t_{ebq} = 8 \times 2ns + 50ns + 8 \times 2ns + 50ns = 132ns$

4 (2 puntos) Calcule el tiempo medio de acceso a este sistema de memoria teniendo en cuenta las siguientes características y estimaciones de la memoria cache unificada:

- tiempo de acceso: 2 ns.
- política de lectura: *Out of Order Fetch.*
- política de escritura: aplazada (CBWA).
- *Hit ratio*: 90 %
- al reemplazar un bloque, el 25 % de las veces dicho bloque se encuentra modificado.

Considerando un comportamiento similar en las lecturas y en las escrituras, y teniendo en cuenta que el único mecanismo de que dispone la cache para reducir el tiempo de penalización es lectura *Out of Order Fetch.* El tiempo medio de acceso se calcula como sigue teniendo en cuenta el tiempo de escritura de un bloque y de lectura de la primera palabra calculados en el apartado anterior:

$$\begin{aligned} \bar{t}_{acc} &= t_{acc_{cache}} + (1 - Hr) \times t_{penal} = t_{acc_{cache}} + (1 - Hr) \times (t_{lp} + P_{modif} \times t_{ebq}) = \\ &= 2ns + (1 - 0.9) \times (52ns + 0.25 \times 132ns) = 10.5ns \end{aligned}$$

5 (1 punto) A este computador se le dota de memoria virtual paginada, con direcciones virtuales de 32 bits. Para realizar la traducción se utilizan 2 niveles de tablas de páginas. Las páginas son de 4 KB, cada entrada de cualquier tabla de páginas ocupa una palabra y cada tabla de traducción ocupa una página. Indique y justifique el formato de las direcciones virtuales.

Como cada página tiene 4 KB (2^{12}), el campo desplazamiento tiene 12 bits. Según se dice cada tabla de traducción ocupa una página y cada entrada una palabra. Es decir, como son todas de igual tamaño, los campos de índice de primer nivel (Zona) y de segundo nivel (Página) también.

Así que las direcciones virtuales tienen la siguiente interpretación:

Zona	Página	Desplazamiento
10 bits	10 bits	12 bits

Nótese que cada tabla de página tiene 1.024 entradas de una palabra y por lo tanto ocupa 4 KB. Es decir, una página como dice el enunciado.

6 (1 punto) Para agilizar los accesos se dota al computador de caches separadas (instrucciones y datos) con las mismas características de la unificada y de sendas TLBs con tiempo de acceso de 2 ns. Justifique si es posible solapar la traducción en TLB con el acceso a la memoria caché.

Para poder solapar el acceso a las TLBs con el acceso a las memorias cache, es necesario que la parte de la dirección virtual que no se traduce ("Desplazamiento") sea suficiente para seleccionar el conjunto en la cache y el byte dentro del bloque correspondiente.

En este caso no se pueden solapar dichos accesos porque el campo "Desplazamiento" tiene 12 bits que no son suficientes para seleccionar el conjunto (8 bits) y el byte (6 bits).

7 (2 puntos) Calcule el tiempo medio de acceso a este sistema de memoria teniendo en cuenta las siguientes características y estimaciones:

- Memoria caches
  - tiempo de acceso: 2 ns.
  - política de lectura: *Out of Order Fetch.*
  - *Hit ratio*: 90 %
  - Memoria cache de datos: política de escritura aplazada (CBWA).
  - Memoria cache de datos: al reemplazar un bloque, el 25 % de las veces dicho bloque se encuentra modificado.
- Memoria virtual
  - tiempo de acceso a las TLBs: 2ns.
  - *Hit ratio*: 95 %
  - no hay fallos de páginas

Como no se pueden solapar los accesos a TLBs y memorias cache, el tiempo medio de acceso será la suma del tiempo medio de traducción y el tiempo medio de acceso al sistema de memoria cache y memoria principal. Nótese que habrá que calcular este tiempo para acceso a datos y a instrucciones ya que en el caso de la cache de instrucciones los bloques no se modifican.

Para calcular el tiempo medio de traducción hay que tener en cuenta que, en caso de fallo en TLB, hay que traducir con dos niveles de tablas de página en las que cada entrada ocupa una palabra. Así que se usará el tiempo de lectura de una palabra calculado en el apartado 3.

$$\bar{t}_{trad} = t_{acc_{TLB}} + (1 - Hr) \times t_{penal} = t_{acc_{TLB}} + (1 - Hr) \times (2 \times t_{lp}) = 2ns + (1 - 0.95) \times (2 \times 52ns) = 7.2ns$$

El tiempo medio de acceso a la cache de datos es el calculado en el apartado 4. Así el tiempo medio de acceso a datos es:

$$\bar{t}_{acc_D} = 7.2ns + 10.5ns = 17.7ns$$

El tiempo medio de acceso a la cache de instrucciones se calcula de forma similar pero teniendo en cuenta que la probabilidad de que el bloque a sustituir esté modificado es 0:

$$\bar{t}_{cache_I} = t_{cache_I} + (1 - Hr) \times t_{penal_I} = t_{acc_{cache}} + (1 - Hr) \times t_{lp} = 2ns + (1 - 0.9) \times 52ns = 7,2ns$$

Resultando un tiempo medio de acceso a instrucciones:

$$\bar{t}_{acc_I} = 7.2ns + 7.2ns = 14.4ns$$

NOTAS: No determinada **DURACIÓN: 90 minutos**
REVISIÓN: No determinada **PUNTUACIÓN: 10 puntos**